Paralelismo Híbrido Aplicado a Soluciones de Problemas de Datos Masivos

Maria A. Murazzo[*], Maria Fabiana Piccoli[#], Nelson R. Rodriguez[*], Diego Medel[*], Jorge N. Mercado[&], Federico Sanchez[**], Ana Laura Molina[***], Martin Tello[**]
[*]Departamento de Informática – FCEFy N, UNSJ.
[#]Departamento de Informática – FCFMy N, UNSL.
[&]Departamento de Matemática – FI, UNSJ.
[**]Alumno Avanzado de la Carrera Licenciatura en Ciencias de la Computación.
[***]Alumno Avanzado de la Carrera Licenciatura en Sistemas de Información.

marite@unsj-cuim.edu.ar, mpiccoli@unsl.edu.ar, nelson@iinfo.unsj.edu.ar, mdiego88@gmail.com, jorgenmp@gmail.com, fedegsancheza@gmail.com, lauramolina@outlook.com, martinl.tello@gmail.com

Resumen

Con el uso masivo de Internet, se está en presencia de un fenómeno donde la aceleración tanto del crecimiento del volumen de datos capturados y almacenados, como la creciente variación en los tipos de datos, hace que las técnicas tradicionales para el procesamiento, análisis y obtención de información útil deban ser redefinidas para formular nuevas metodologías.

Este trabajo aborda las líneas de investigación relacionadas con el objetivo de definir técnicas o metodologías computacionales para mejorar tanto los tiempos de respuesta y la escalabilidad de los sistemas desarrollados, como así también solucionar los inconvenientes presentes en las soluciones existentes. Esto posibilita la transferencia de los logros y experiencias adquiridos, permitiendo, además, verificar la aplicabilidad de los métodos o técnicas desarrollados en problemas reales con uso de datos masivos.

Palabras clave: Datos Masivos, Computación de Alto Desempeño, Arquitecturas Multiprocesadores, Sistemas Distribuidas.

Contexto

El presente trabajo se encuadra dentro del área de I/D “Procesamiento Distribuido y Paralelo” y en particular dentro del proyecto de investigación "Evaluación de arquitecturas distribuidas de commodity basadas en software libre", código E1038, el cual ha sido aprobado en la última convocatoria de CICITCA (duración dos años y unidad ejecutora al Departamento de Informática de la FCEFyN de la UNSJ).

Introducción

Trabajar *con grandes volúmenes de datos,* implica un gran desafío debido a la necesidad de explorar un universo de nuevas tecnologías, las cuales no sólo hacen posible la obtención y procesamiento de los datos sino también realizan su gestión en un tiempo razonable, lo que permite contar con una fuente inagotable de problemas para aplicar de técnicas computacionales de alto desempeño.

El crecimiento de la cantidad de datos es algo cotidiano y obedece a la proliferación de diferentes fuentes de generación de información, como son la web, aplicaciones de imagen y vídeo, redes sociales, dispositivos móviles, sensores, Internet de las cosas, etc., todas ellas capaces de generar según IBM [1], más de 2.5 quintillones de bytes diarios. Este

aumento en la cantidad de datos demanda nuevas estrategias para su almacenamiento, procesamiento y análisis, conllevando a un cambio de paradigma en las arquitecturas de cómputo, los algoritmos y los mecanismos de procesamiento.

Ejemplos cotidianos de datos masivos son el número de imágenes subidas diariamente a las redes sociales (300 millones en Facebook, 45 millones en Instagram), los videos vistos por día en YouTube (2 billones), la cantidad de mensajes de texto enviados, la cantidad mensual de búsquedas en Twitter, el tráfico mundial en Internet, entre otros [2]. Esto no sólo es aplicable a las actividades desarrolladas diariamente en Internet, sino también en aquellas relacionadas a fenómenos naturales como el clima o datos sismográficos, entornos referidos a la salud, bioinformática, seguridad o al ámbito empresarial.

Además de la gestión del volumen de datos, gran parte de la información requerida para la toma de decisiones y la resolución de problemas de índole general proviene de información no estructurada, almacenada o accedida no necesariamente en estructuras clásicas de almacenamiento como matrices, registros de bases de datos, etc.

Restringirse al uso de información estructurada lleva, muchas veces, a representar una visión parcial del problema y dejar fuera de consideración información de gran importancia para la resolución efectiva del mismo.

Frente a esta problemática se ha popularizado el término Big Data [3], el cual es usado para describir grandes conjuntos de datos, que exhiben las propiedades de variedad, volumen, velocidad, variabilidad, valor y complejidad. Hablar de Big Data es hacer referencia a datos multidimensionales, estructurados o no estructurados.

Big data es un área de investigación focalizada en recolectar, examinar y procesar grandes conjuntos de datos con el objeto de descubrir patrones, correlaciones y extraer información de ellos [4] . Por lo general, esta tarea se implementa mediante el uso de diferentes técnicas, entre ellas se encuentra machine learning supervisado y no supervisado[5], técnicas computacionalmente muy costosas, ya sea en la fase de aprendizaje o en la de predicción llegando a ser intratables de manera secuencial cuando involucra grandes volúmenes de datos [6].

Estos aspectos hacen que los sistemas de cómputo convencionales sean muchas veces inapropiados para lograr un procesamiento adecuado, por lo que una alternativa llega a considerar técnicas de computación de alta prestaciones (HPC) con el fin de aumentar la velocidad de procesamiento [7,8].

Generalmente, HPC es referenciado como una evolución de los sistemas de cómputo convencionales, los cuales permiten realizar operaciones de cómputo intensivo y mejorar la velocidad de procesamiento. HPC involucra diferentes tecnologías tal como los sistemas distribuidos y los sistemas paralelos; incluyendo a los cluster de computadoras, cloud computing, tarjetas gráficas y computadoras masivamente paralelas. Todos estos entornos son ideales para resolver aplicaciones científicas, computacionalmente costosas con manejo de grandes cantidades de datos, a fin de lograr resultados en menor tiempo.

La constante demanda de mayores prestaciones hizo que la industria de los procesadores se encontrara en una situación límite respecto al cumplimiento de la ya conocida Ley de Moore sobre rendimiento del hardware [9]. La evolución de los sistemas de computación con multiprocesadores ha seguido dos líneas de desarrollo: las arquitecturas multi-core (multi-núcleos) y las arquitecturas many-cores (muchos-núcleos o muchos-cores). En el primer caso, los avances se centraron en el desarrollo de mejoras con el objetivo de acelerar las aplicaciones, generalmente secuenciales, con por ejemplo la incorporación de varios núcleos de procesamiento.

La industria tomó la idea de las supercomputadoras existentes e incorporó más procesadores a sus desarrollos, surgiendo así computadoras con 2, 3, 4, 8 o más

procesadores por unidad central (multi-core). En el caso de las arquitecturas many-cores, los desarrollos se centraron en optimizar el desempeño de aplicaciones paralelas. Dentro de este tipo de arquitectura se encuentran las tarjetas gráficas o GPUs [10]. La característica mas relevante de este tipo de arquitectura es la capacidad de ofrecer cores simples y rápidos con acceso a una jerarquía de memoria compartida.

Sin embargo, los sistemas de memoria compartida tienen como inconveniente principal la cantidad de memoria disponible, la cual es limitada. Una alternativa a estas arquitecturas son las arquitecturas con memoria distribuida, las cuales permiten incrementar el espacio de almacenamiento (principal y secundario) aunque deben pagar el precio de la latencia de la red para llevar a cabo las comunicaciones [11]. Ejemplos de sistemas con memoria distribuida son los cluster y cloud computing [12]. Estos sistemas permiten conectar un gran número de máquinas (nodos) y utilizar la memoria perteneciente a diferentes procesadores, logrando eliminar el problema de la limitación de memoria RAM.

Una tercera alternativa son los sistemas híbrido, los cuales permiten combinar las características de ambos sistemas: Memoria compartida y Memoria Distribuida, e incrementar la capacidad y poder de cómputo de los sistemas computacionales. Esto posibilita la ejecución en paralelo de múltiples procesos y threads con distintas administraciones de memoria.

La presente propuesta tiene como objetivo desarrollar y aplicar técnicas computacionales híbridas de alto desempeño para la adquisición, tratamiento y análisis de datos masivos en ambientes mixtos de computación.

Líneas de Investigación, Desarrollo e Innovación

En función de lo explicado anteriormente, la linea de investigación pretende el desarrollo de herramientas de software adecuadas para resolver problemas de datos masivos en ambiente paralelos híbridos. Dichas herramientas tendrán como objetivo mejorar el desempeño de cada una de las etapas involucradas en la solución de este tipo de problemas: adquisición, análisis y visualización de los resultados.

El desempeño de cada una de las técnicas y/o herramientas propuestas será evaluado tanto en las soluciones computacionales a los problemas específicos planteados como así también en problemas de datos masivos reales. Para ello se pretende realizar un análisis respecto a:

1. *Rendimiento:* En este caso no solo se consideraran las métricas de rendimiento estándares como aceleración, eficiencia y costo, sino también plantear otras, las cuales estarán relacionadas a las características de los ambientes de computación híbridos, por ejemplo referidas a la distribución y asignación de trabajo, y a los factores limitantes como el desempeño de la red subyacente, las comunicaciones, sincronizaciones y gránulo de la computación.
2. *Calidad:* En este caso, la evaluación se centrará en las propiedades de los resultados obtenidos, es su Relevancia y Precisión.

Además se analizarán otros factores como la portabilidad (tanto de ambiente de computación como de problema), escalabilidad y robustez de cada una de las técnicas desarrolladas.

Resultados y Objetivos

Como objetivo de la investigación se ha planteado el análisis, diseño e implementación de soluciones computacionales eficientes a problemas de datos masivos mediante la aplicación de modelos de programación y técnicas de Computación de Alto Desempeño en ambientes híbridos.

Los ambientes de computación híbridos están formados por arquitecturas multiprocesador (multi-core y many-core) y

arquitecturas distribuidas (clusters y cloud) como así también distintos modelos de memoria. Trabajar en ambientes híbridos, permite aplicar una estrategia de paralelización más efectiva mediante múltiples niveles de paralelismo y reducción del overhead de comunicación. Esto es importante cuando se trabaja con grandes volúmenes de datos debido a la necesidad de contar con una arquitectura escalable.

Por ello, es necesario investigar sobre:

- Arquitecturas Híbridas: Características básicas, adaptabilidad a problemas de diferente naturaleza.
- Modelos de programación estándares e híbridos existentes.
- Problemas de Datos Masivos: Características, etapas involucradas, aspectos paralelos de cada una.
- Herramientas existentes para resolver problemas de datos masivos: análisis de sus características, ventajas y desventajas, factibilidad de aplicación en ambientes computacionales híbridos.
- Problemas reales con uso de datos masivos.
- Análisis de desempeño en ambientes homogéneos y heterogéneos, parámetros de evaluación.

En particular con respecto a los datos masivos es necesario:

- Analizar las soluciones existentes para cada una de las etapas involucradas en problemas de datos masivos, evaluando sus limitaciones e inconvenientes de uso en ambientes híbridos.
- Analizar las características asociadas a resolver problemas con información estructurada y no estructurada.
- Diseñar e implementar nuevas técnicas para las etapas de adquisición, tratamiento y análisis de datos masivos a fin de proveer de una/s herramienta/s para resolver problemas reales aplicando técnicas de computación de alto desempeño en ambientes computacionales híbridos.
- Analizar y elaborar métricas de rendimiento para evaluar el desempeño de los desarrollos en ambientes de computación híbridos.

Para realizar las investigaciones de esta linea, el equipamiento necesario estará en diferentes ubicaciones, en la UNSJ se cuenta con un cluster multi-core y en la UNSL se cuenta con una cluster de características híbridas, un cluster con varios multi-cores y many-cores. Además se cuenta con acceso a equipamiento perteneciente a diferentes unidades académicas y a centros de computación de alto desempeño como aquellos nucleados en el Sistema Nacional de Computación de Alto Desempeño dependiente del Ministerio de Ciencia, Tecnología e Innovación de la Nación.

Formación de Recursos Humanos

Los resultados esperados respecto a la formación de recursos humanos son hasta el momento el desarrollo de 1 tesis doctoral y 2 tesis de maestría. Respecto a las carreras de grado, se están ejecutando 4 tesinas de grado.

Además se prevé la divulgación de los temas investigados, tanto a través del dictado de cursos de postgrado/actualización, como de publicaciones en diferentes congresos y revistas del ámbito nacional e internacional.

Referencias

[1] "¿Qué es Big Data?" [Online]. Available: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/.

[2] "Blog Cisco Cansac » Tráfico de datos móviles crecerá casi 10 veces en los próximos cinco años, predice estudio Cisco Visual Networking Índex (VNI)." [Online]. Available: http://gblogs.cisco.com/cansac/trafico-de-datos-moviles-crecera-casi-10-veces-en-los-proximos-cinco-anos-predice-estudio-cisco-visual-networking-index-vni/?

doing_wp_cron=1473174616.1210250854492187500000.

[3] A. Mcafee, E. Brynjolfsson, and H. Org, “Big Data: The Management Revolution SPOTLIGHT ON BIG DATA,” 2012.

[4] Y. Zhai, Y.-S. Ong, and I. W. Tsang, “The Emerging "Big Dimensionality",” *IEEE Comput. Intell. Mag.*, vol. 9, no. 3, pp. 14–26, Aug. 2014.

[5] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, “Data mining with big data,” *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[6] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. 2016.

[7] Y. You, S. L. Song, H. Fu, A. Marquez, M. M. Dehnavi, K. Barker, K. W. Cameron, A. P. Randles, and G. Yang, “MIC-SVM: Designing a Highly Efficient Support Vector Machine for Advanced Modern Multi-core and Many-Core Architectures,” in *2014 IEEE 28th International Parallel and Distributed Processing Symposium*, 2014, pp. 809–818.

[8] M. Alexander, W. Gardner, B. Wilkinson, M. J. Sottile, T. G. Mattson, C. E. Rasmussen, Y. Robert, and F. Vivien, “Introduction to High Performance Computing for Scientists and Engineers Chapman & Hall/CRC Computational Science Series.”

[9] C. A. Mack, “Fifty Years of Moore’s Law,” *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 2, pp. 202–207, May 2011.

[10] W. Hwu, K. Keutzer, and T. G. Mattson, “The Concurrency Challenge,” *IEEE Des. Test Comput.*, vol. 25, no. 4, pp. 312–320, Jul. 2008.

[11] N. Miranda, “Cálculo en tiempo real de identificadores robustos para objetos multimedia mediante una arquitectura paralela CPU-GPU,” 2016.

[12] K. Kaur and A. K. Rai, “A Comparative Analysis: Grid, Cluster and Cloud Computing,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 3, no. 3, pp. 2278–1021, 2014.